Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 10:35:35
REFERENCE NO: 227

Author Names & Affiliations

- Jonathan Fresnedo-Ramírez - Ohio Agricultural Research and Development Center, Department of Horticulture and Crop Science, The Ohio State University.

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Plant Sciences, plant genetics, plant breeding and applied bioinformatics

Title of Submission

Current and future context of cyberinfrastructure for applied bioinformatics for plant sciences

Abstract (maximum ~200 words).

Research in plant sciences has a high rate of data generation, and “omics-based” data generation will continue to accelerate for the next 30 years. Bioinformatics and computational biology methods will be main tools for the analysis of that data, and such methods have to be upon a solid cyberinfrastructure. Emerging fields in plant sciences include: pangenomic studies, haploid-resolved genomics, genomics of introgression, synthetic biology of autotroph organisms, protein engineering, gene regulation, cell differentiation control, the understanding of ageing and molecular farming. Infrastructure for the advancement of these fields is probably already available in terms of hardware; however, gaps need to be closed in terms of software, which has to be optimized and intuitive considering basic, translational and applied science. Contained computing and abstraction for efficient data analyses, as well as training to do so, will become main demands by plant scientists working in translational science such crop improvement. Little attention has been paid to cyberterrorism and the vulnerability of data generated. Other neglected topics include data centralization and duplicity, modularity for the continuous improvement of cyberinfrastructure at the institutional level while making it environmentally friendly, sustainability and resiliency overtime and space, and ensuring the development of the human resources required.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

In plant sciences, the exponential curve of generation of “omics” (i.e. genomics, phenomics, transcriptomics. proteomics, metabolomics and epigenomics) will continue for a while (some researchers say another 30 years). The use of this data varies enormously. There continues to be a challenge to appropriate manage of data from the generation of raw data to the transfer, distribution, sharing and final storage. It has to allow subsequent reproducibility and data integration (https://doi.org//10.1186/1752-0509-8-S2-I1). The application of the data is also

very diverse and in some cases the storage of data may not need to be long term, but in other cases, it may be.
In living organisms such as plants (also applicable to fungi and other microorganisms) researchers are pursuing what I call “giving birth their organism’s specific Laplace’s demon”. Researchers pursue the development of models in which everything about the research target is known. It means that researchers want to pursue “at least” what has been achieved in reference species (i.e. Arabidopsis thaliana or Saccharomyces cerevisae), and be able to produce large databases having all information around that organism.
In plant science, a topic that is going to become really common is the issue of pangenomic studies (https://doi.org/10.1093/bib/bbw089), at intracellular (each plant cell contains three genomes: nuclear, mitochondrial and plastid), intrapespecies, and multispecies complexes, including organ-specific biomes. Certainly, the number of genomes that can be sampled without so much regulatory drag in the plant kingdom will empower researchers to generate data that easily may surpass data generated for humans (just to give an example). It has begun with Arabidopsis and Maize, in which thousands of accessions are being sequenced, and processed in both, independent assembly and joint assembly, with the purpose of better understanding the genomic makeup of a most comprehensive representation of a species and its intraspecies “-omic” variation. This topic requires the development of new types of files (VCF, BAM, SAM and HDF5 are not enough anymore) which may represent multiple features of accessions, and more likely considering multiple dimensions (a genetic variant with a specific physical position in a specific accession, in a specific time point, with a resemblance in multiple other accessions or clusters of related accessions, in which the same variant is located in a distinct physical position but with certain modifications, which are not necessarily genomic). Here improvements and generation of intuitive compression algorithms are still needed, which would help to the accessibility.
Haploid-resolved genomes (i.e. haploid-resolved diploid and haploid-resolved polyploid genome sequences) will become a major pathway to understanding genetic configuration of relevant traits regarding evolution, conservation and breeding for agriculture. This has been done for human beings already (https://doi.org/10.1038/nbt.3200) and its implementation in plant sciences is going to be very relevant since it will allow a better understanding of the conformation of diplotypes and multi-plotypes in a diverse range of species, and define how alleles interact for the exhibition of relevant traits. It will require intensive targeted sequencing of not only selected accessions but also their pedigree to better resolve phases, assembly and haplotype resolved through software.
Understanding of the genomic modifications triggered by introgression of traits (and therefore genome segments) from related species is a major pending task that plant researchers are expecting to tackle as sequencing and algorithms become available (https://doi.org/10.1371/journal.pgen.1003477). Several commercially relevant traits such as biotic and abiotic stresses are being recovered from crop related wild species, and such related species usually have a distinct biological history (e.g. mating system) in comparison to the domesticated crops. The introgression of traits and the linkage drag associated can be examined to understand the extent of the genomic changes are undergoing in the crop-host genome.
Synthetic biology of complex organisms in larger scale is going to be another big topic with protein engineering for molecular farming. The manipulation of metabolism for the study/obtaining of specific materials (e.g. biopolymers) for which synthetic versions in laboratories cannot duplicate the characteristics of the original version. This requires the generation of modeling multi-domain proteins and large assemblies and to consider those materials as a collection of dynamic elements (RNA 3D structure prediction, RNA 3D modeling, movement, transcription, protein synthesis, protein folding, biomolecule dynamics, etc.) for which the recording of data has to be very intensive and readily available for analysis and in which Artificial Intelligence is expected to contribute to the identification of patterns and description of processes and its integration with systems biology (https://doi.org/10.1093/bioinformatics/btu769).
Active fields from which additional needs will be realized include systems biology, gene regulation, cell differentiation control and the understanding of ageing. Plenty of these studies have been approached in some extent in humans, but in other biological systems we are just beginning to explore and define strategies for data generation. Given the plasticity and manipulability of other organisms, the potential to generate data is practically limitless. Quantum physics and chemistry of plants will grow as a field since research has yielded answers that will allow the manipulation and engineering of systems related to energy production and storage.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

I consider that in terms of advance computing and computing optimization, a good job is being performed, we have more capable hardware at this time and seems that great advances are taking place (we have computers with terabytes of RAM and GPU processing seems promising, for example). One thing I consider is lacking is the development of software that allows us to take full advantage of the hardware and then address the questions that we have. Much of the data going to be generated from biological samples for which phenotypic data is available will rely on using multilevel modeling, which is mainly pursued through iterative methods which need to evolve to be adaptive

according to the data being analyzed (such as the likelihood function). Software to perform these tasks taking advantage of multi-parallel and cloud computing for scientific analysis is just starting; however, this task is little supported and rewarded, and not as attractive as development of commercial software. Algorithms may already exist to take advantage of hardware potential but the implementation represents an important gap which will require (quoting Leroy Hood) “a cross-disciplinary environment composed of biologists, chemists, computer scientists, engineers, mathematicians, physicists, and physicians speaking common discipline languages” in addition to strong financial and occupational support for human resources enduring in such development and implementation.
There are examples of tools such as the language and framework R (https://cran.r-project.org), which (although is not the favorite language of hardcore programmers) it is supported, maintained and used for a diverse group of scientists and engineers. The evolution of such tools as shown by the Julia language (https://julialang.org) show hope for the development and implementation of tools that will optimize the use of the current hardware cyberinfrastructure without endangering the intuitiveness and the ability to adapt current bioinformatics and statistical pipelines.
Open Source hardware and software is a big need in order to keep and nurture community-feeling in research. Emphasis on certain areas of development and improvement are being demanded to be released as open source only in order to standardize some parts of data analysis and then ensure reproducibility as well as keep the safety and integrity of data and results.
Contained computing (encapsulation) and abstraction for efficient data analysis, as showed by the docker hub (e.g. https://hub.docker.com), is a very interesting approach, which along with other types of software containers, will allow to take advantage of computing power to pursue high computing demands for analyses. However, the implementation of these tools coupled with additional machine learning and blockchains to ensure the integrity of information generated and shared among teams around the globe is in its infancy in applications for plant sciences. The Integrated Breeding Platform(https://www.integratedbreeding.net) through CyVerse is an example of the application of part of these technologies for some of the most important crops worldwide. The results will suggest whether this approach is appropriate for the collaboration and use of genomic data and the application of results in a field of high relevance for humanity: improvement of crops for the production of food and other raw materials. Deployment of cyberinfrastructure and intensive training is needed in order to successfully use the information already generated and information that is going to be generated.
Applications and related knowledge generated through the analysis of the information (i.e. data repositories) may be a target of new types of cyberterrorism (unseen now but latent), in which deletion or ransom of information will be a risk for research institutions and private industry.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

One aspect to consider is about centralization of data. Plenty of institutions are moving to centralize computing power and datacenters but a main concern is the safety of the data in eventual adverse conditions, and how data may be duplicated and where. The safety of the data refers not only to the integrity of the device containing the data and its functionality, but also to the way in which data is stored and may be inspected and manipulated by external agents. Certainly, a large concentration of data and power would allow high scale management and maintenance but a low ability for customization.
Some other institutions are relegating the development of datacenters and facilities to their specific research groups. This model demands modularity (some other calling it composable) for growing. It means the datacenter is going to grow depending on the needs of the group, for which usually faculty members are going to contribute to add cores or storage units. Then technology (some people demanding it as open source) that allows the scalability of small datacenters with very particular needs and aims (some talk of a minecraft/lego for datacenters and bioinformatics cores) is necessary. This is appropriate for researchers not only working in smaller institutions or in a concentrated field of study, it also useful for entrepreneurship efforts and subsequent product development. It may also be appropriate for the definition of scientific roles that a researcher or a team will be playing. Thus, basic science will have specific needs, which are not necessarily shared with translational science, and finally with engineering and product development. Allowing a certain degree of customization but support for an upper level of cyberinfrastructure will be beneficial for several emerging and exploratory research topics.
An issue that I have noted in my career is that researchers and institution are paying little attention to make our cyberinfrastructure sustainable and resilient. Little is being done in order to decrease the footprint in terms of carbon, heat and radiation generation, noise management, material recycling (not only plastic and metal, also gases), and in general self-sufficiency among datacenters and bioinformatics facilities (https://doi.org/10.1038/nclimate1786). No research institutions make announcements of environmental friendly/committed constructions of data centers and facilities, in order to follow the Power Usage Effectiveness of the federal Data Center Optimization Initiative (DCOI). Most of the effort comes from the private sector and interaction with universities will be needed. Can we do more with the heat that is being generated? What degrees of self-sufficiency may be achieved in datacenters and bioinformatics facilities?
Another aspect to regard is the development of human resources for running facilities, for providing “customer service”. Computer Science

is strong but the development of talents in all topics related to bioinformatics is little encouraged. Undergraduate courses and specializations are few in graduate school, for both theoretical and applied perspective, and are even more scarce in undergraduate education institutions. Today we are tending to hire undergraduate IT people with little understanding of the characteristics of data generated in bioinformatics and the needs of data management. On the other hand, several of the current bioinformatics facility managers are self-taught, these are colleagues that started either as biologist/chemists and learned coding or programmers that got used to the coding and data manipulation needed for the biological mining of information.

Consent Statement